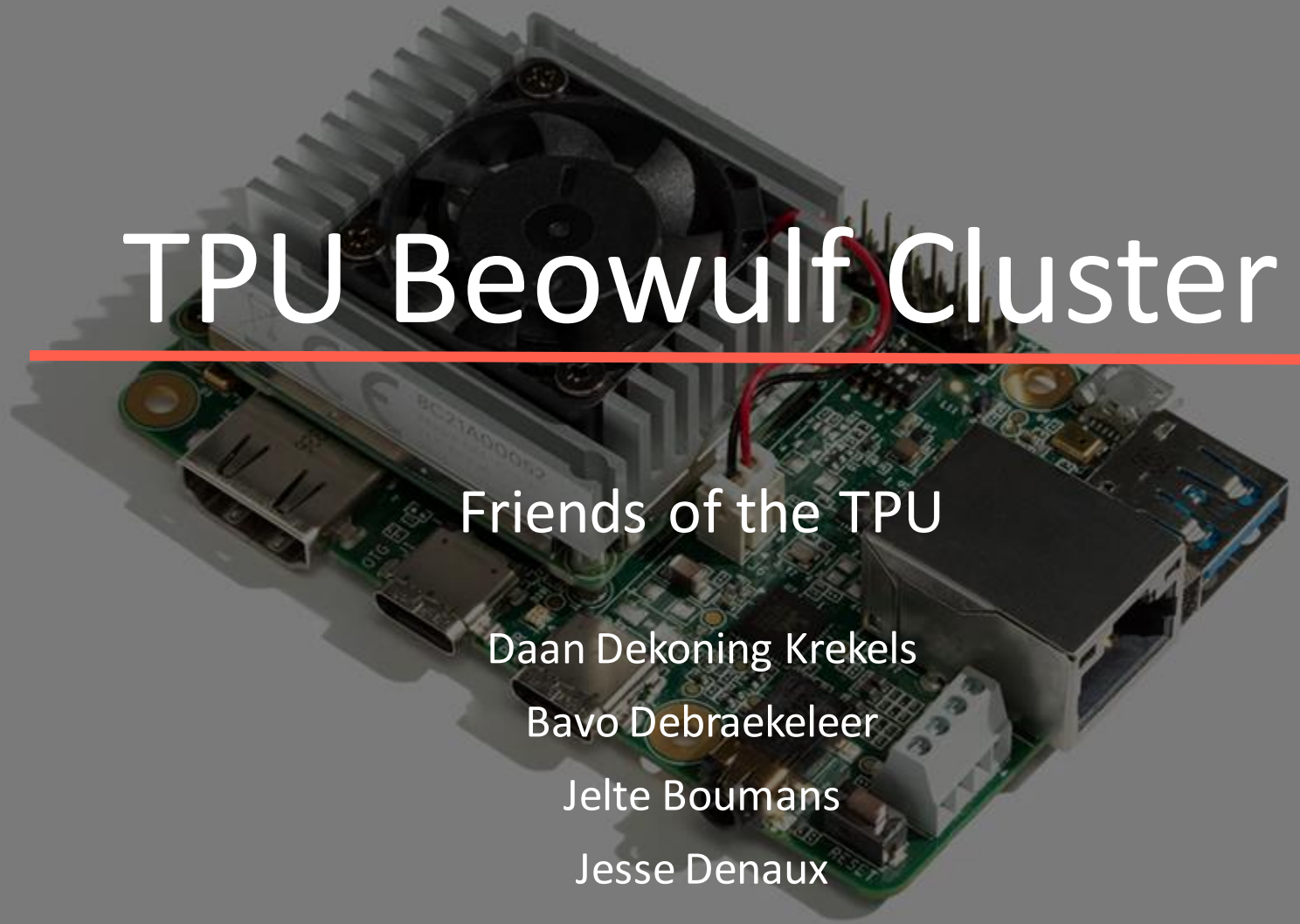
TPU Beowulf Cluster



Friends of the TPU

Daan Dekoning Krekels

Bavo Debraekeleer

Jelte Boumans

Jesse Denaux

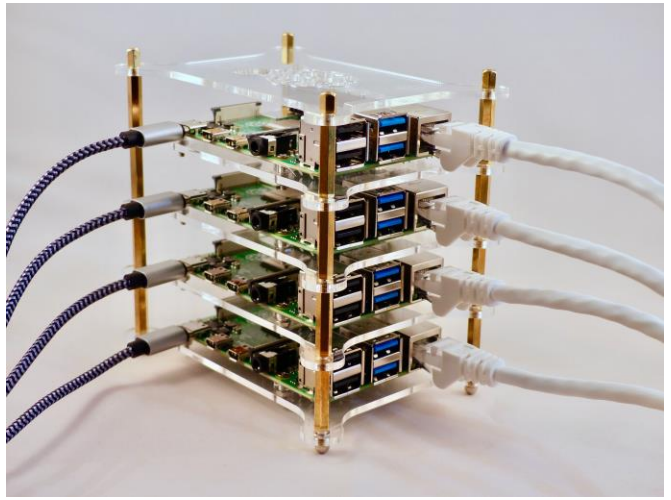
Tarun Singh

Inhoudstabel

- Doelstelling — *Tarun*
- Coral Dev Board — *Bavo*
- Mendel Linux — *Bavo*
- TensorFlow (Lite) model — *Tarun*
- Aanpak Cluster — *Daan & Tarun*
- Behuizing — *Jelte*
- Alternatieven — *Daan*
- Conclusie — *Daan*

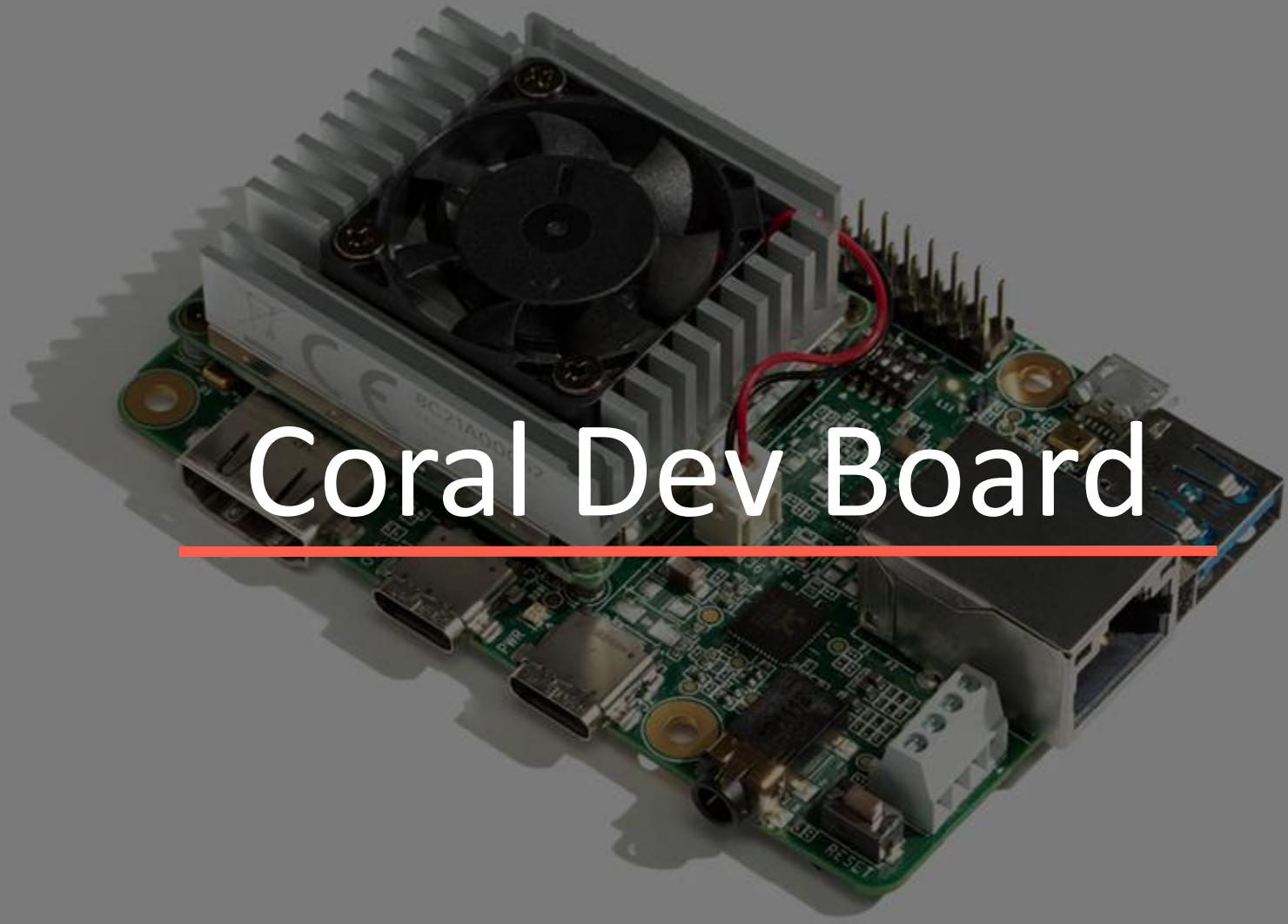
Doelstelling

Doelstelling



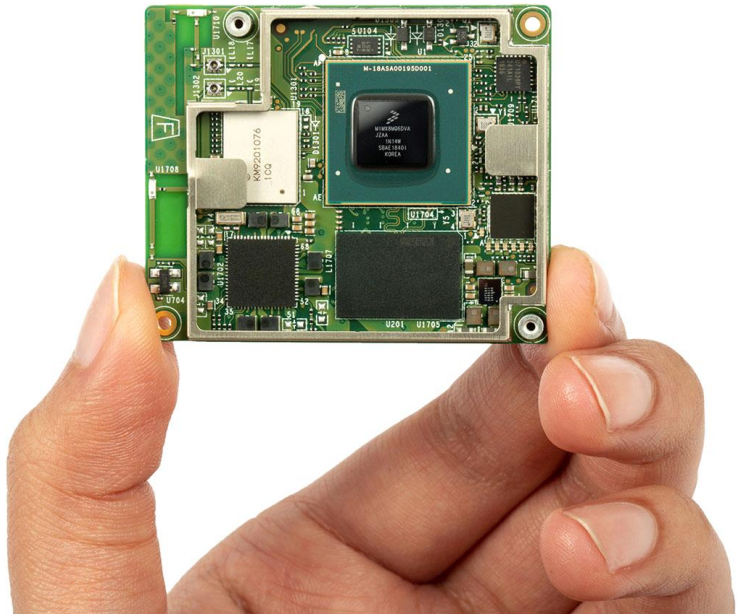
- Cluster met TPU’s bouwen
- Private netwerk
- Snelheid verhogen door de rekenkracht te verdelen
- NFS en Open MPI
- Runnen van een AI model op de cluster

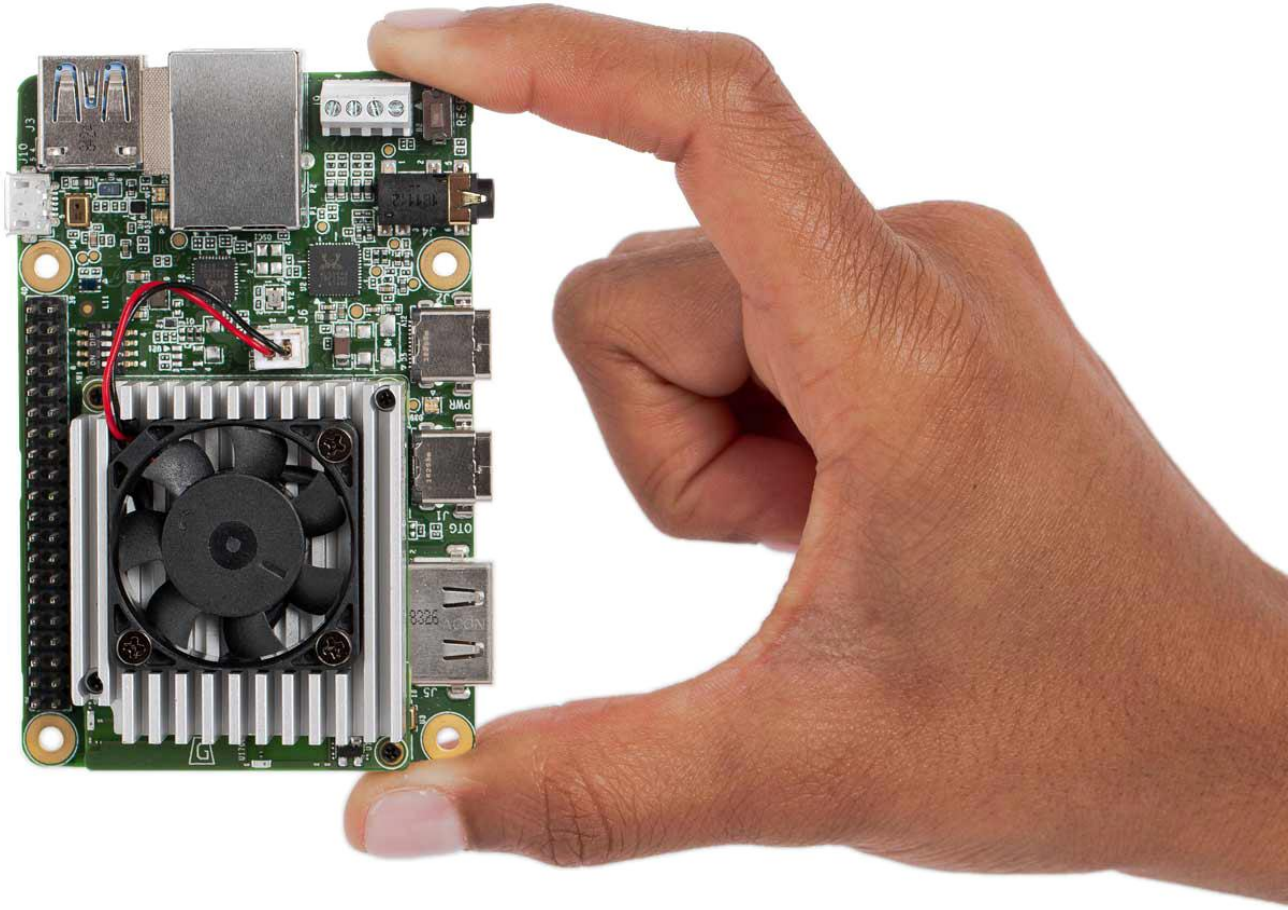
Coral Dev Board



Single Board Computer

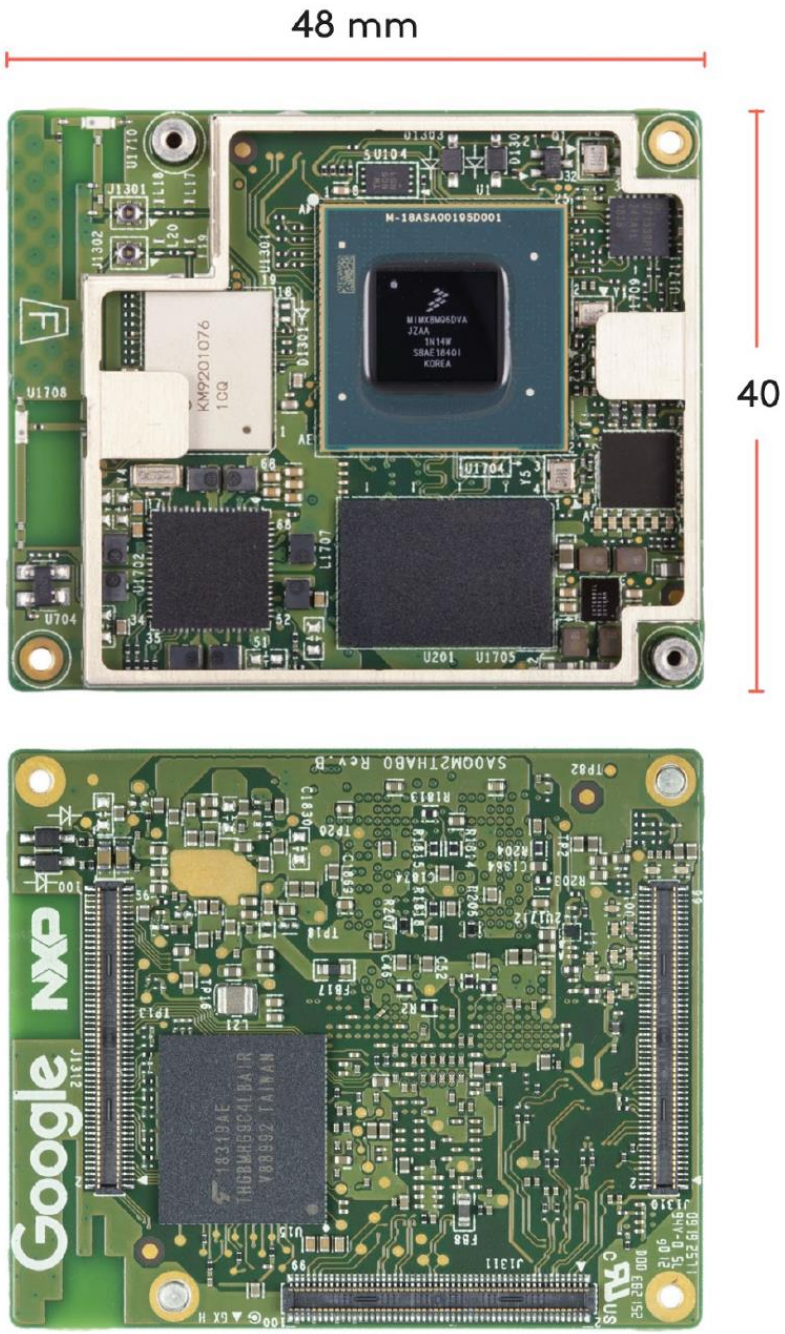
- Edge TPU System-on-Module (SoM)
- Baseboard met connectoren

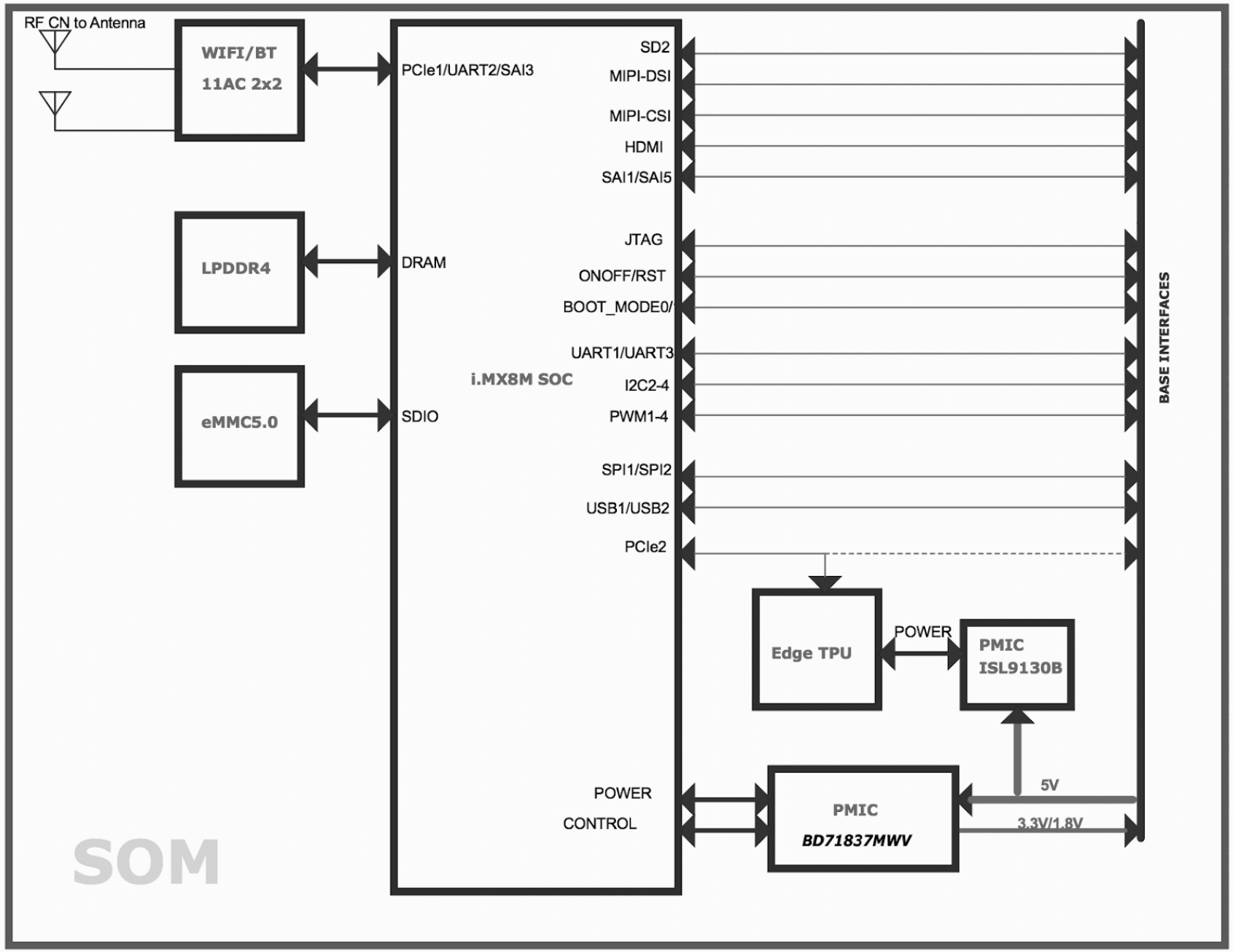




Edge TPU SoM

- NXP i.MX 8M SoC (MIMX8MQ6DVA)
  - Cortex-A53/M4F
  - Vivante GC7000 Lite GPU en VPU
- Google Edge TPU ML accelerator coprocessor
- PMIC (BD71837MWV)
- Microchip ATECC608A Cryptographic coprocessor
- Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5 GHz)
- Bluetooth 4.2
- 8/16 GB eMMC
- 1/4 GB LPDDR4

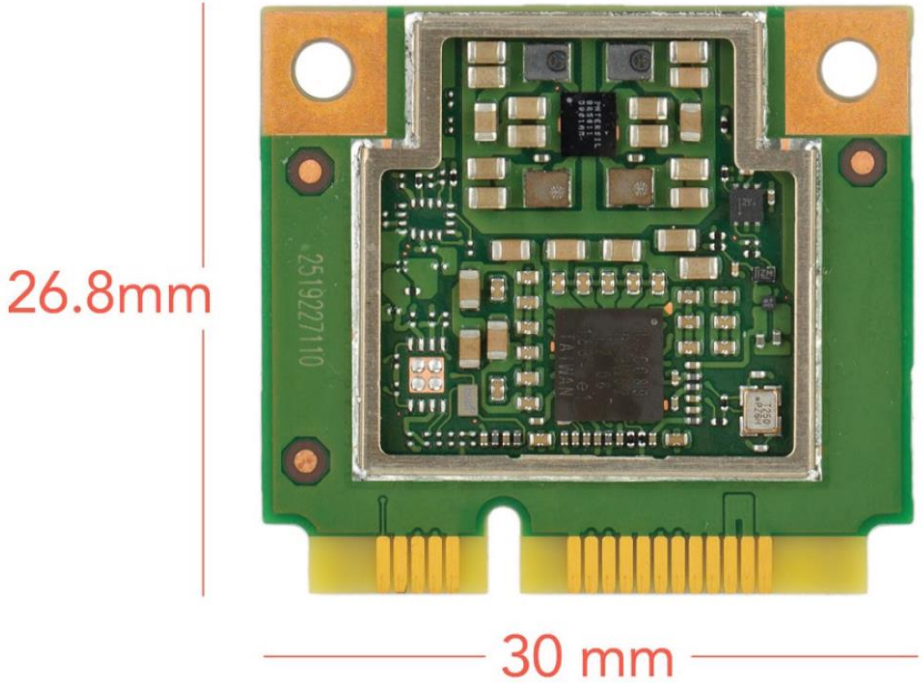


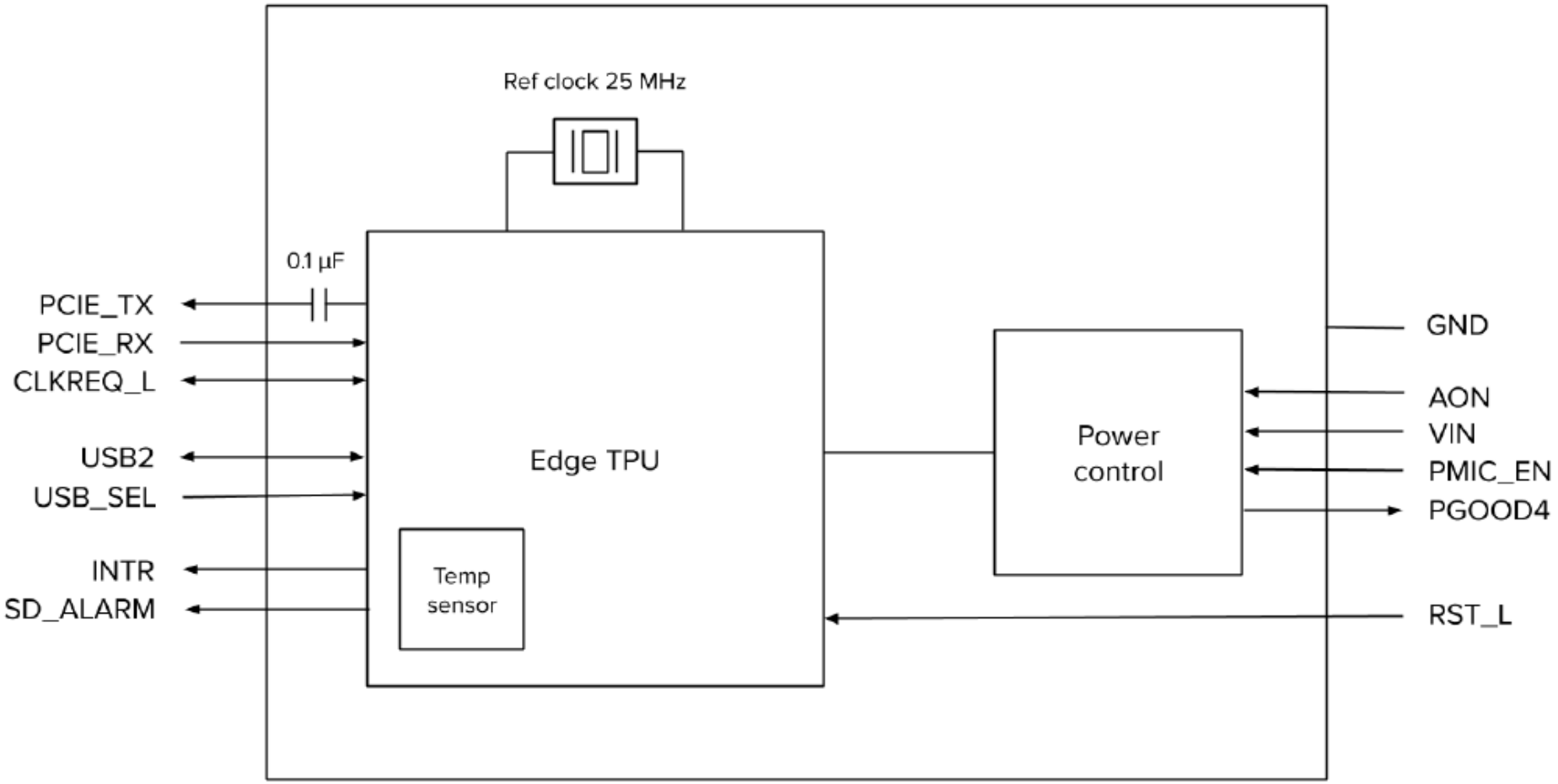

RF CN to Antenna
WIFI/BT
11AC 2x2
LPDDR4
eMMC5.0
PCIe1/UART2/SAI3
DRAM
SDIO
i.MX8M SOC
SD2
MIPI-DSI
MIPI-CSI
HDMI
SAI1/SAI5
JTAG
ONOFF/RST
BOOT_MODE0/
UART1/UART3
I2C2-4
PWM1-4
SPI1/SPI2
USB1/USB2
PCIe2
BASE INTERFACES
Edge TPU
POWER
PMIC
ISL9130B
5V
3.3V/1.8V
POWER
CONTROL
PMIC
BD71837MWV
SOM

Google Edge TPU ML accelerator

- Tensor Processing Unit (TPU)
  - Google ASIC voor ML
  - Hoog volume laagprecisie-berekeningen
  - Laag energie verbruik

- Google Edge TPU ML accelerator
  - Lokaal TensorFlow Lite modelen
  - 4 TOPS @ 2 W
  - PMIC ISL9130B
  - PCIe Gen2 x1 interface

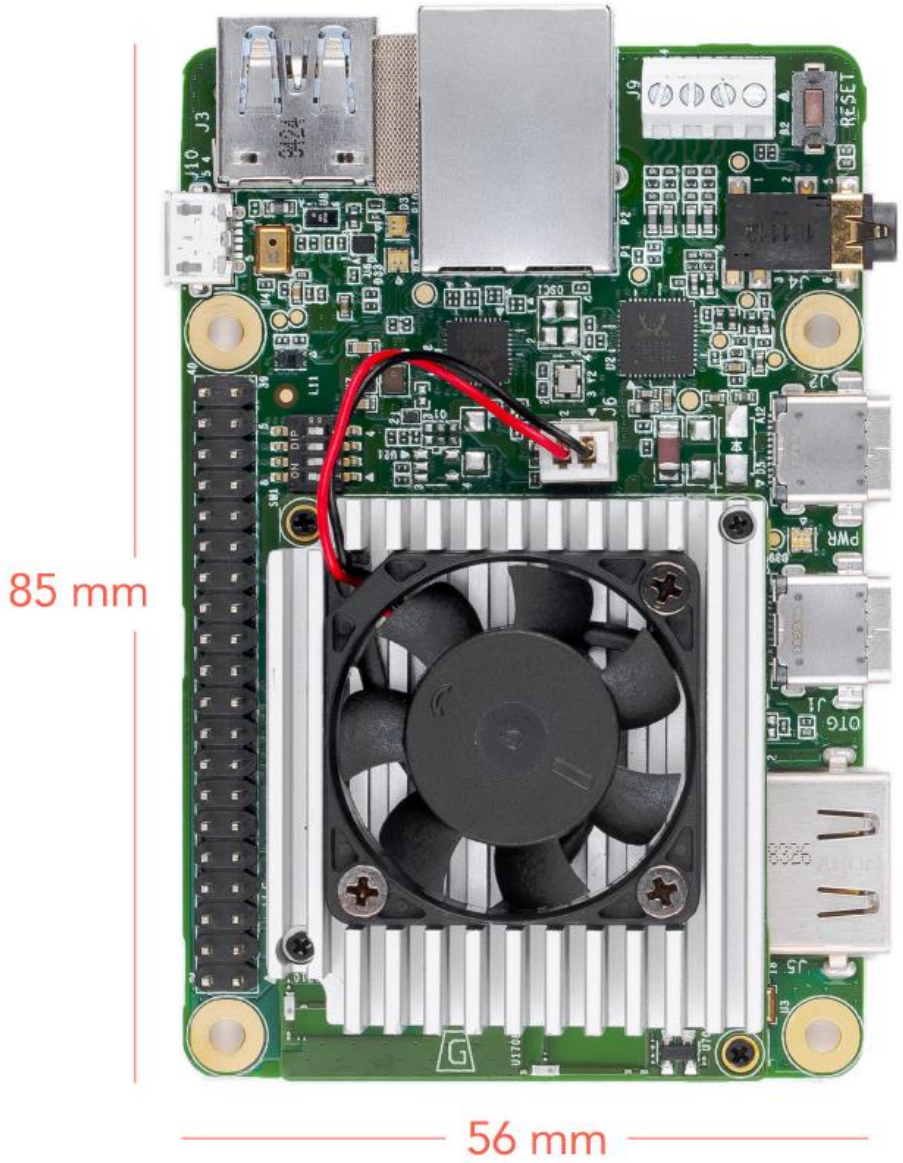


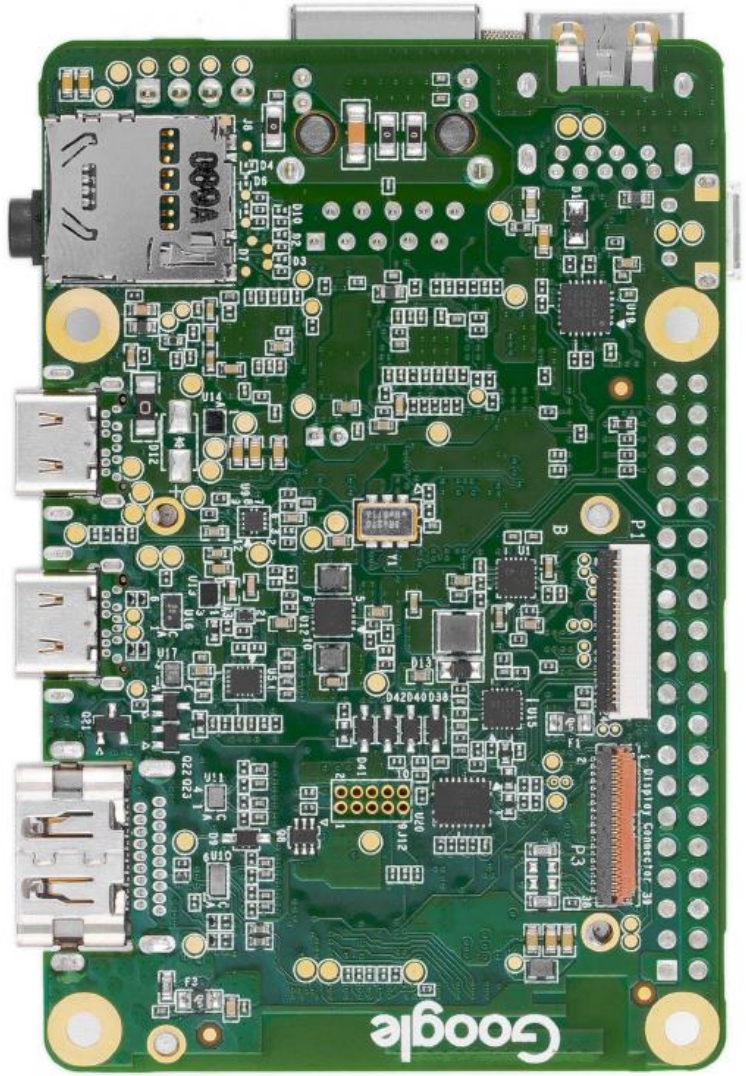


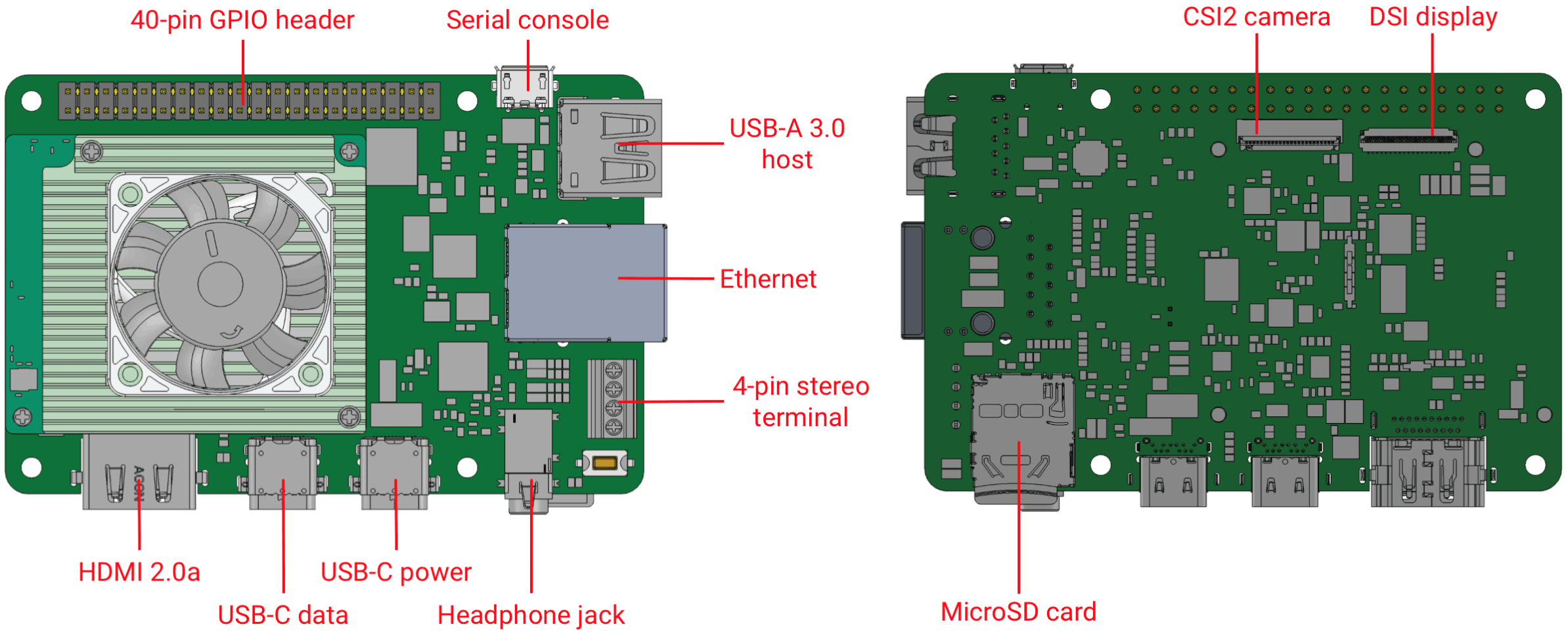
Google's TPUv3 has 8 cores — 4 chips each with 2 cores — with each core having 2 MXUs.

Baseboard

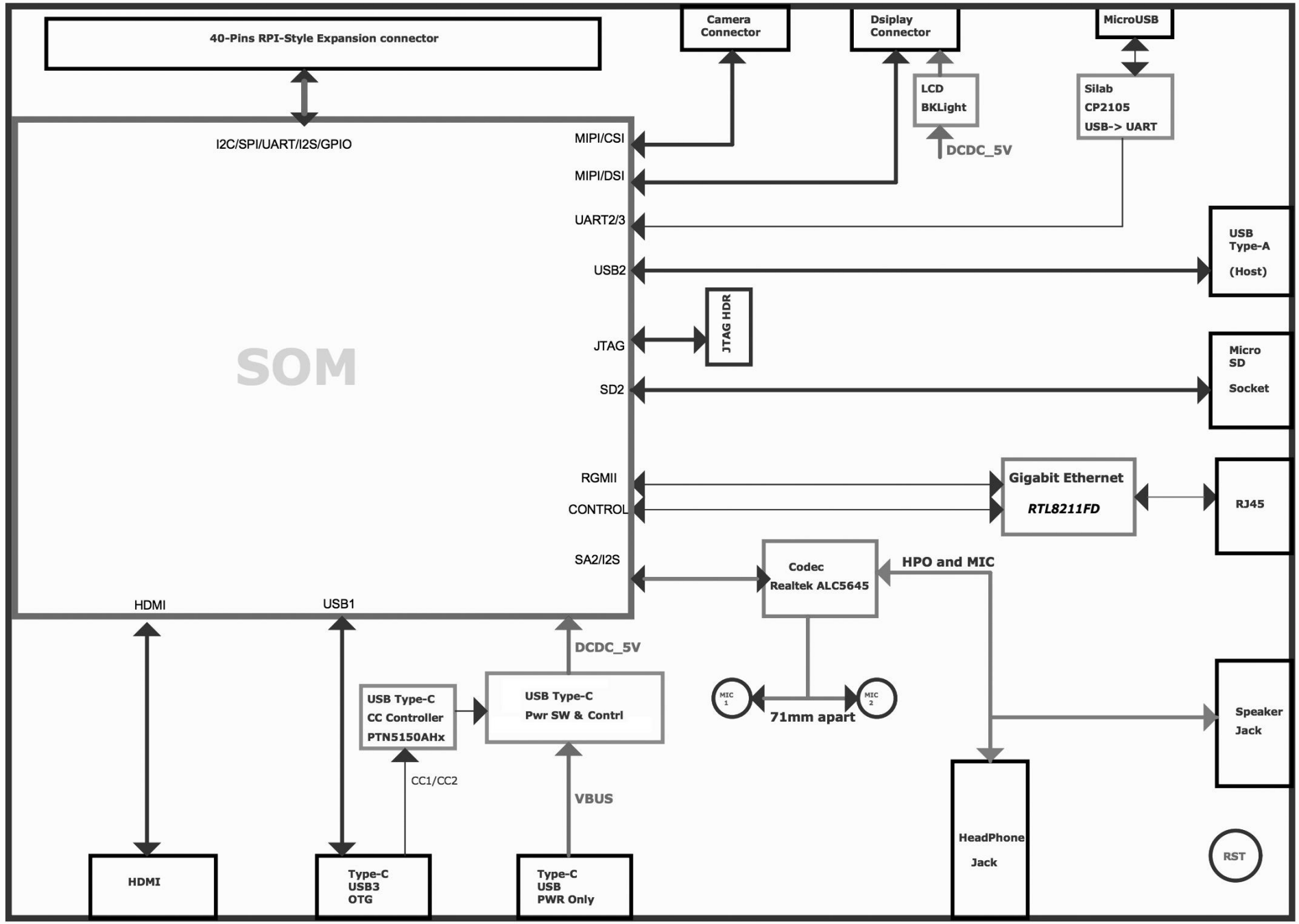
- Prototype board
- Connectiviteit

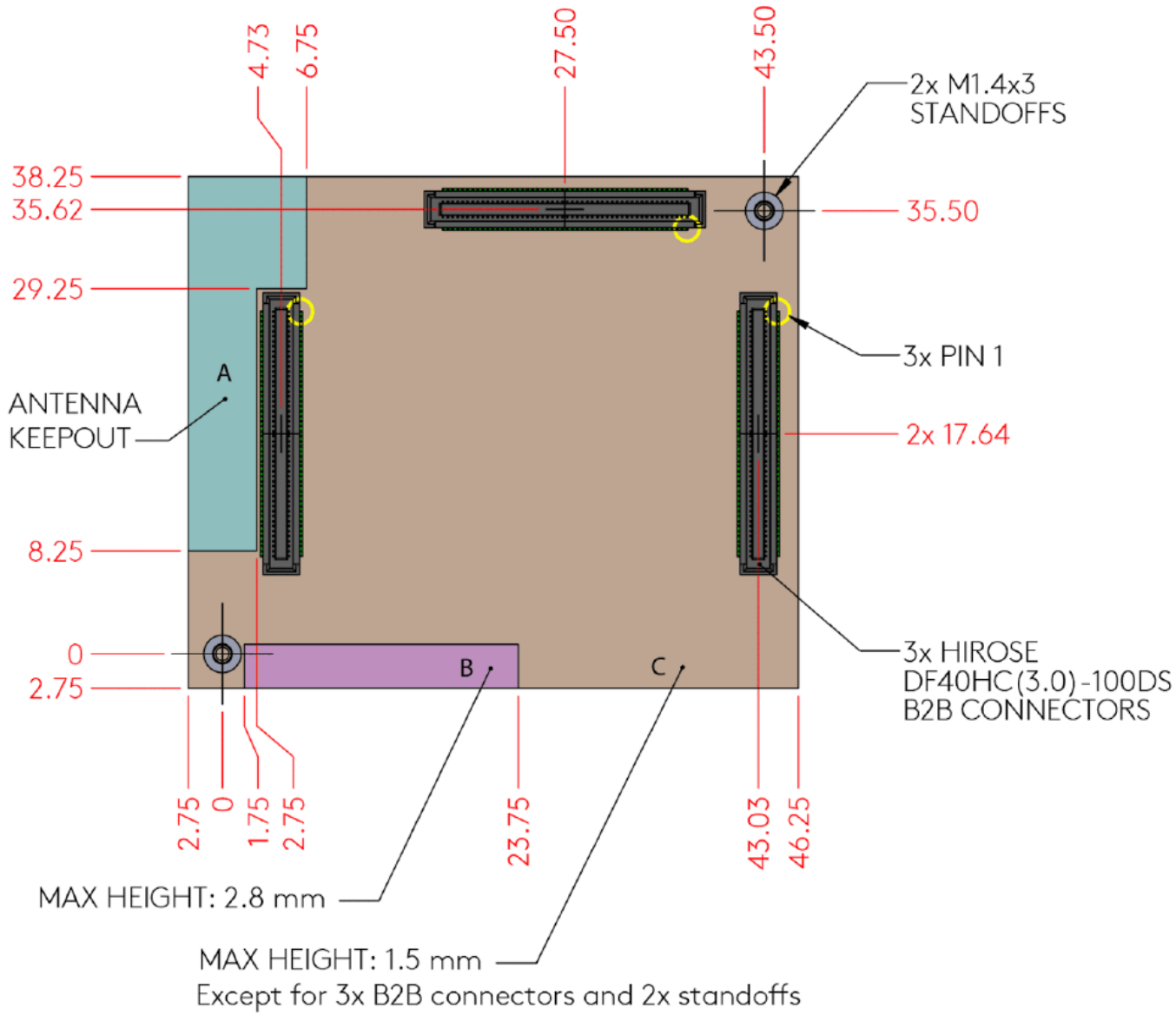


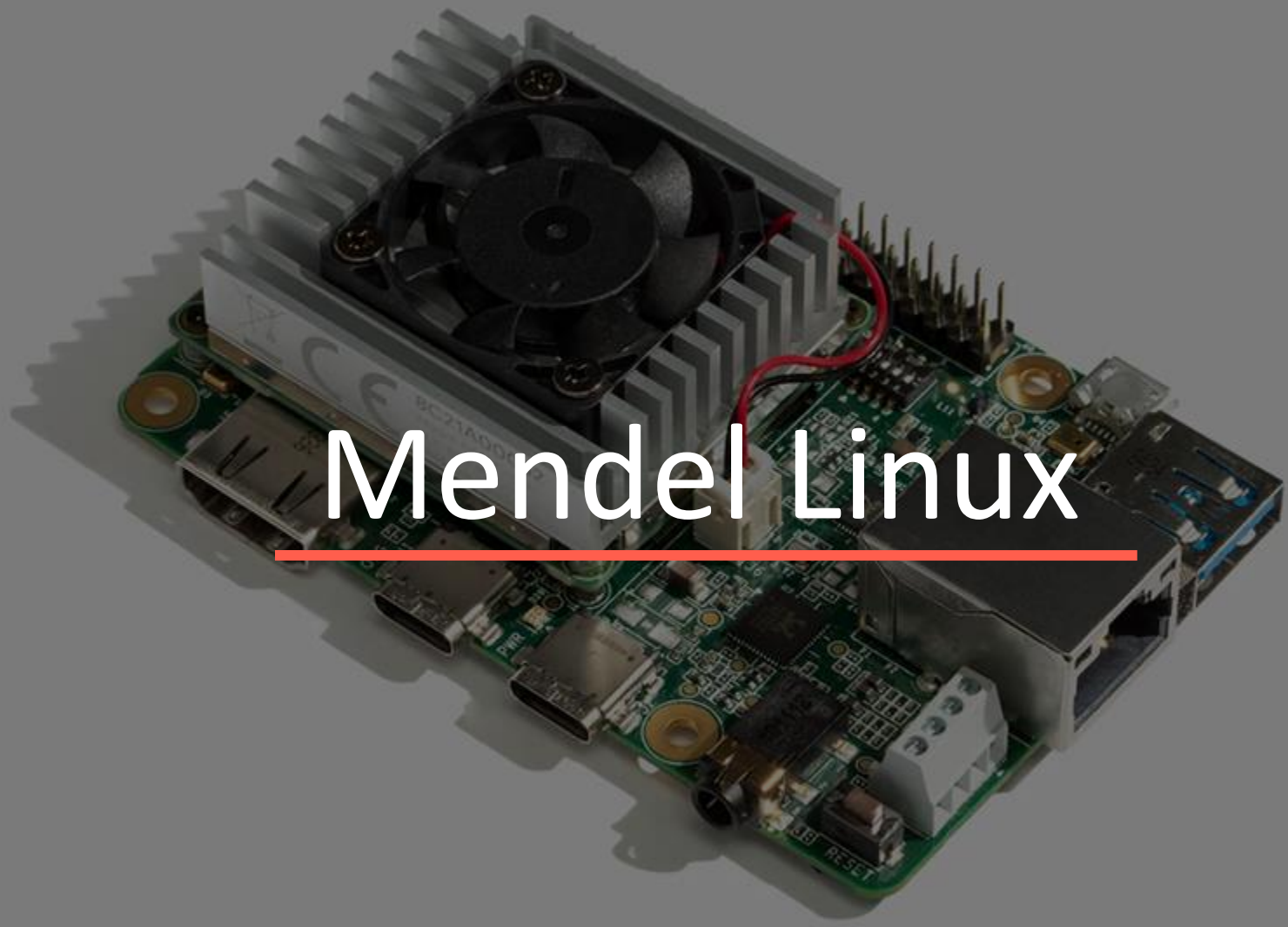



40-pin GPIO header
Serial console
USB-A 3.0 host
Ethernet
4-pin stereo terminal
HDMI 2.0a
USB-C data
USB-C power
Headphone jack
CSI2 camera
DSI display
MicroSD card

5V
3.3V
Ground
GPIO
PWM
I2C
UART
SPI
SAI
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
J10
J3


40-Pins RPI-Style Expansion connector
Camera Connector
Dsiplay Connector
MicroUSB
LCD BKLight
DCDC_5V
Silab CP2105 USB-> UART
I2C/SPI/UART/I2S/GPIO
MIPI/CSI
MIPI/DSI
UART2/3
USB2
USB Type-A (Host)
JTAG HDR
JTAG
SOM
Micro SD Socket
SD2
RGMII
CONTROL
Gigabit Ethernet RTL8211FD
RJ45
SA2/I2S
Codec Realtek ALC5645
HPO and MIC
HDMI
USB1
DCDC_5V
USB Type-C CC Controller PTN5150AHx
USB Type-C Pwr SW & Contrl
MIC 1
MIC 2
71mm apart
Speaker Jack
CC1/CC2
VBUS
HeadPhone Jack
RST
HDMI
Type-C USB3 OTG
Type-C USB PWR Only


4.73
6.75
27.50
43.50
2x M1.4x3
STANDOFFS
38.25
35.62
35.50
29.25
3x PIN 1
A
ANTENNA
KEEPOUT
2x 17.64
8.25
0
2.75
B
C
3x HIROSE
DF40HC(3.0)-100DS
B2B CONNECTORS
2.75
0
1.75
2.75
23.75
43.03
46.25
MAX HEIGHT: 2.8 mm
MAX HEIGHT: 1.5 mm
Except for 3x B2B connectors and 2x standoffs



Mendel Linux

Wat is Mendel?

- *Lightweight derivative* upstream Debian 10 (buster) Linux distributie
- Voorgeïnstalleerde packages voor ML
  - Python 3.6 - 3.9
  - *python3-pycoral*
  - *libedgetpu1-std*
  - *python3-edgetpuvision*
  - *mendel-minimal*

Wat is Mendel?

- Aanpassing voor specifieke hardware Dev Board:
  - Device Tree Overlays
  - Coral "*apex*" PCIe driver
  - Edge TPU runtime
  - PyCoral library
  - Defconfig:
    - Include USB FTDI driver
    - Add micro k8s / k3s deps
  - i2s:
    - Enable I2S output on 40-pin header
  - rt5645:
    - Add a sysfs-based override for hp-detect

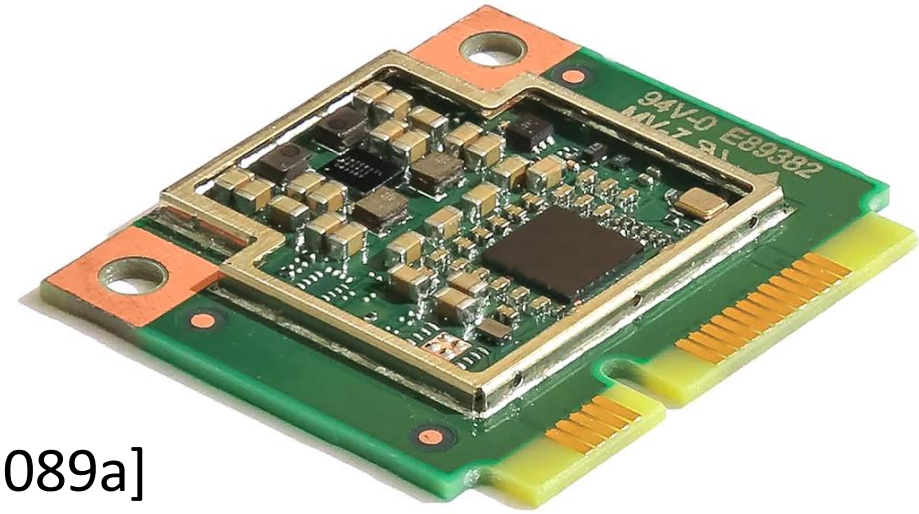
Model Pipelining

- Detecteerd PCI of USB peripherals

```
{'type': 'pci', 'path': '/dev/apex_0'}
```

```
enum edgetpu_device_type {
  EDGETPU_APEX_PCI = 0,
  EDGETPU_APEX_USB = 1,
};

struct edgetpu_device {
  enum edgetpu_device_type type;
  const char* path;
};

struct edgetpu_option {
  const char* name;
  const char* value;
};
```

TPU Detectie



- *lspci -nn | grep 089a*
  - 0001:01:00.0 System peripheral [0880]: Global Unichip Corp. Coral Edge TPU [1ac1:089a]
- */sys/bus/pci/devices/0001:01:00.0*
- */proc/bus/pci/devices*
  - 0100 1ac1089a 1eb 2020000c 0 2010000c 0 0 0 0 4000 0 100000 0 0 0 0 apex

Device Tree Overlay

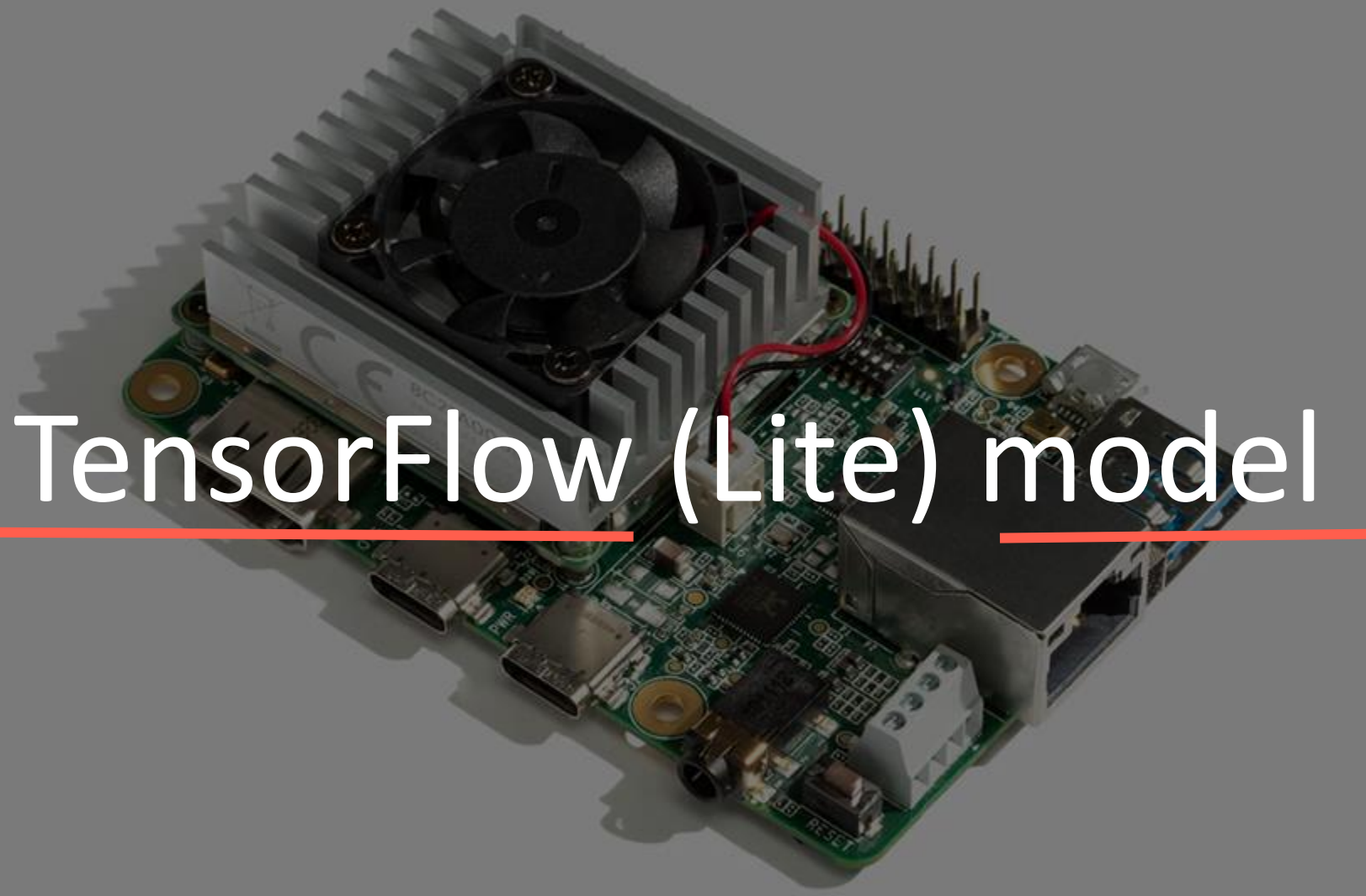
- fsl-imx8mm-columbia.dts
  - *FSL i.MX8MM Columbia board*
  - Onderste laag, 2375 lijnen code
- fsl-imx8mq-phanbell.dts
  - *Freescale i.MX8MQ Phanbell*
  - Toevoeging "*google,edgetpu-audio-card", sound-rt5645*
  - Middelste laag, 2522 lijnen code
- fsl-imx8mq-yorktown.dts
  - Uitbreiding *Freescale i.MX8MQ Phanbell*
  - Toevoeging *Yorktown TPU "bus" device,* met 8x *Akira cores*, bestaande uit 2x *MXU's*
  - Bovenste laag, 2827 lijnen code

Device Tree Overlay

```
imx8mq-phanbell {

        pcie1grp {
                fsl,pins = < 0x230 0x498 0x528 0x12 0x00 0x76 0x13c 0x3a4 0x00 0x05 0x00 0x19 >;
                linux,phandle = < 0x67 >;
                phandle = < 0x67 >;
        };
```

```
pcie1: pcie@0x33c00000 {
        compatible = "fsl,imx8mq-pcie", "snps,dw-pcie";
        reg = <0x0 0x33c00000 0x0 0x400000>, <0x0 0x27f00000 0x0 0x80000>;
        reg-names = "dbi", "config";
        #address-cells = <3>;
        #size-cells = <2>;
        device_type = "pci";
        ranges =  <0x81000000 0 0x00000000 0x0 0x27f80000 0 0x00010000 /* do
               0x82000000 0 0x20000000 0x0 0x20000000 0 0x07f00000>; /* non-
        num-lanes = <1>;
        interrupts = <GIC_SPI 74 IRQ_TYPE_LEVEL_HIGH>,
                <GIC_SPI 80 IRQ_TYPE_LEVEL_HIGH>; /* eDMA */
        interrupt-names = "msi";
        #interrupt-cells = <1>;
```

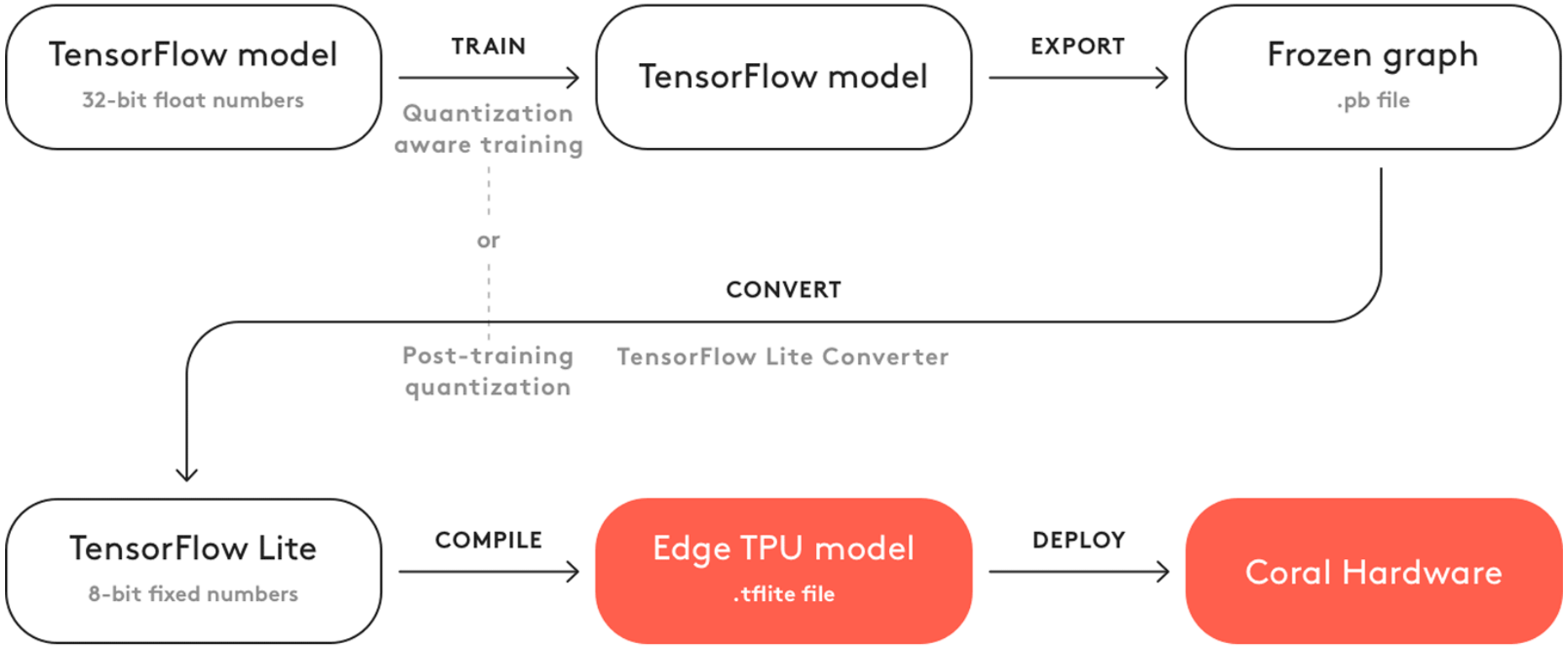
```
yorktown-tpu {
                compatible = "google,yorktown-tpu";

                akira0 {
                        compatible = "google,akira";
                        status = "okay";
                        tpu-gpios = < 0x6c 0x01 0x00 0x6c 0x00 0x01 >;
                };

                akira1 {
                        compatible = "google,akira";
                        status = "okay";
                        tpu-gpios = < 0x6d 0x01 0x00 0x6d 0x00 0x01 >;
                };

                akira2 {
                        compatible = "google,akira";
                        status = "okay";
                        tpu-gpios = < 0x6e 0x01 0x00 0x6e 0x00 0x01 >;
                };

                akira3 {
                        compatible = "google,akira";
                        status = "okay";
                        tpu-gpios = < 0x6f 0x01 0x00 0x6f 0x00 0x01 >;
                };

                akira4 {
                        compatible = "google,akira";
                        status = "okay";
                        tpu-gpios = < 0x70 0x01 0x00 0x70 0x00 0x01 >;
                };

                akira5 {
                        compatible = "google,akira";
```

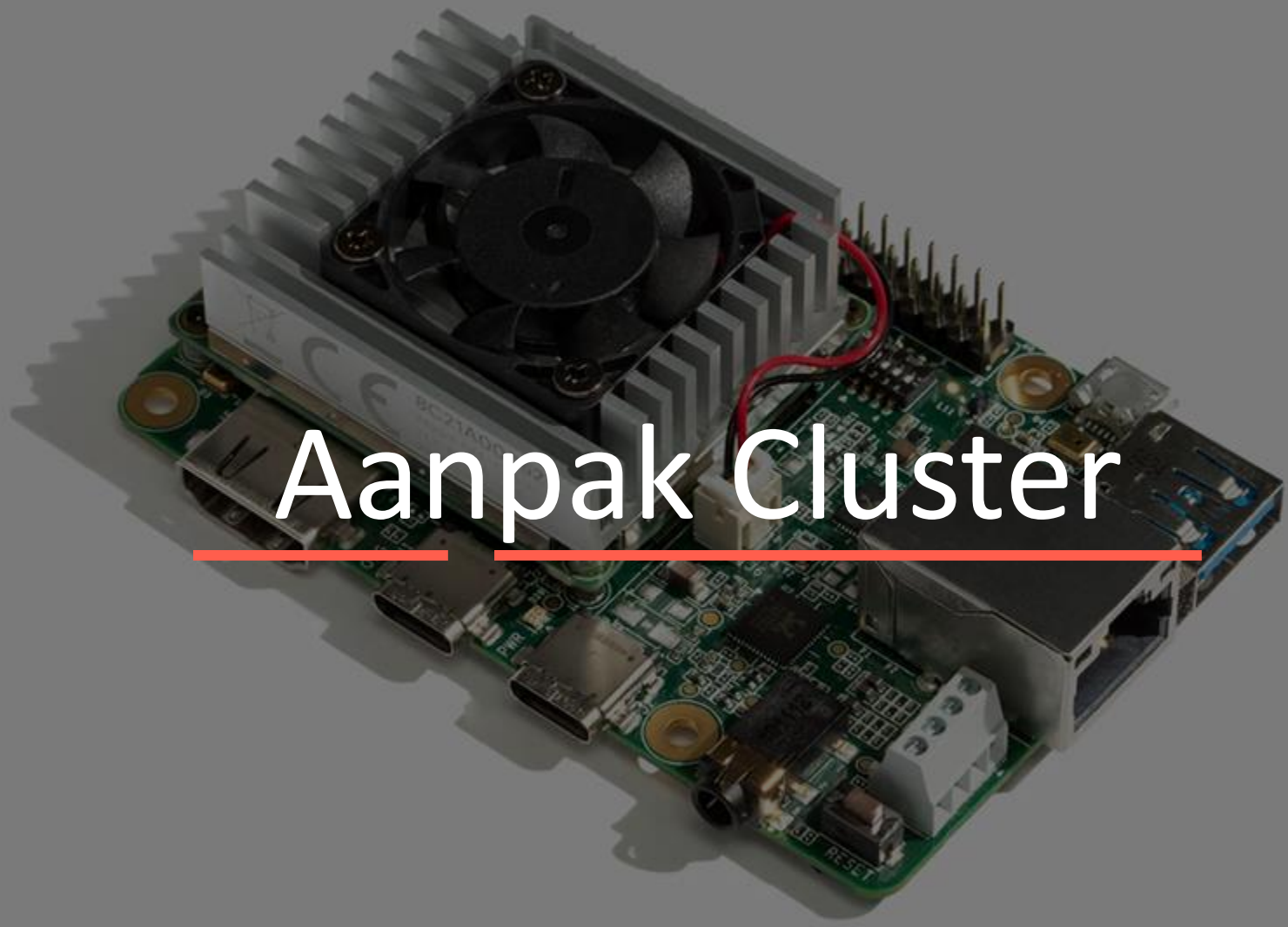


TensorFlow (Lite) model

TensorFlow (Lite) model

- Tensorflow model omzetten naar .tflite bestand (8bit)
- .tflite model compileren

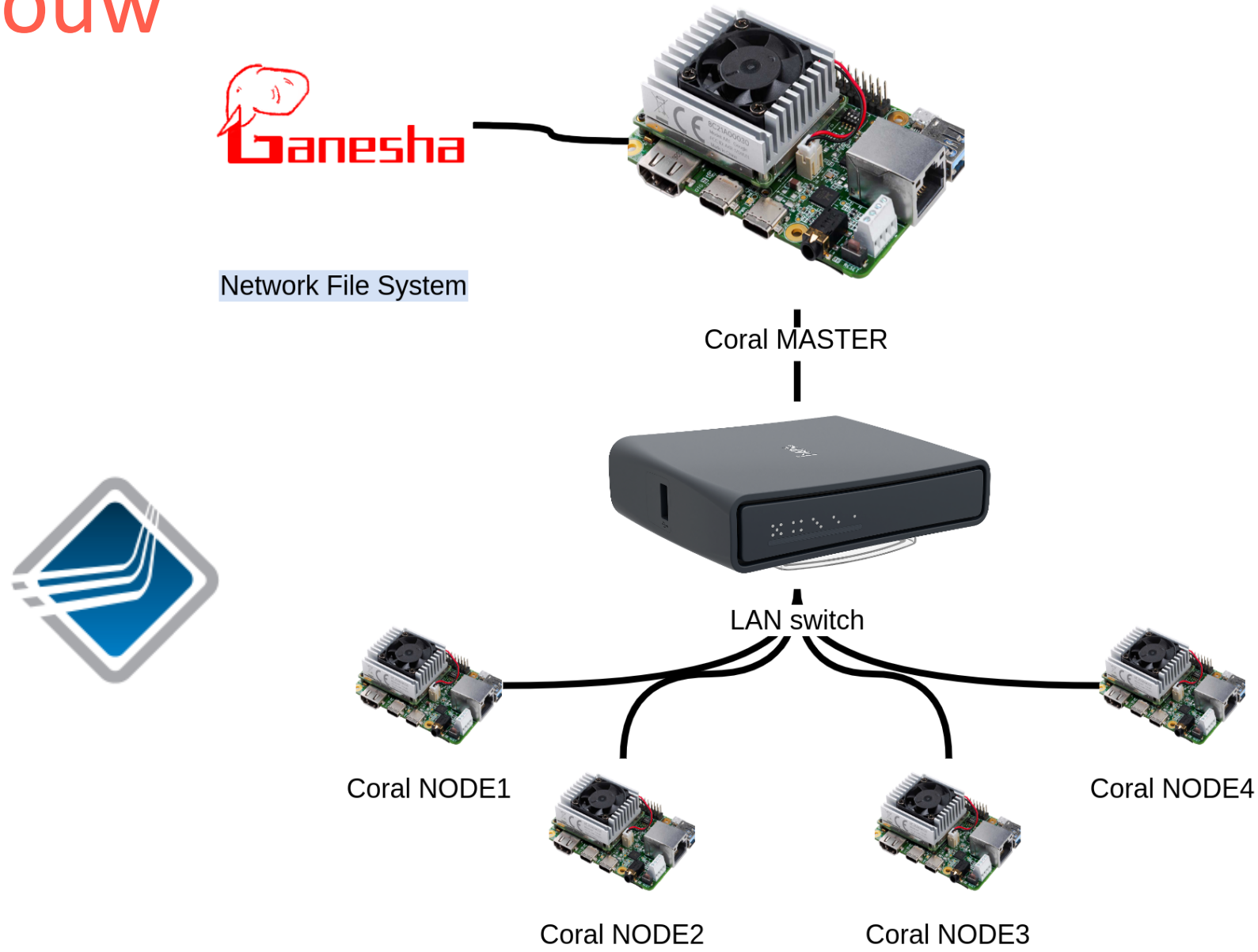




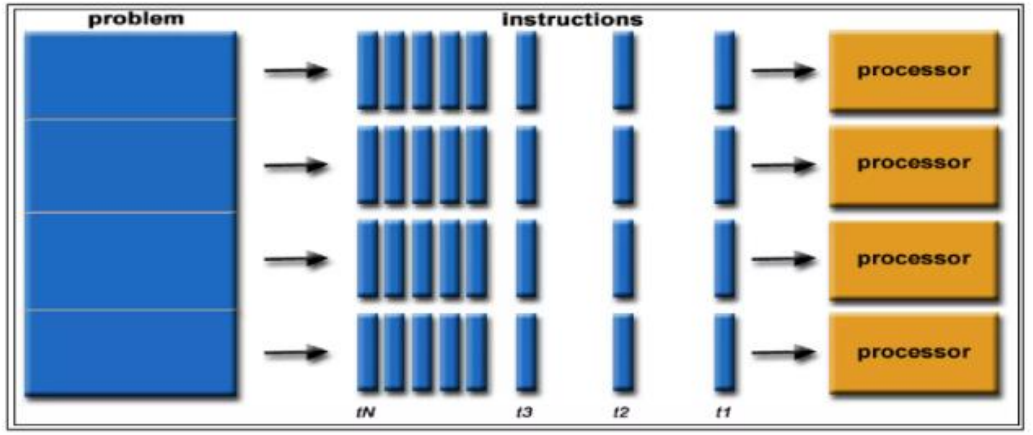
Aanpak Cluster

Algemene opbouw

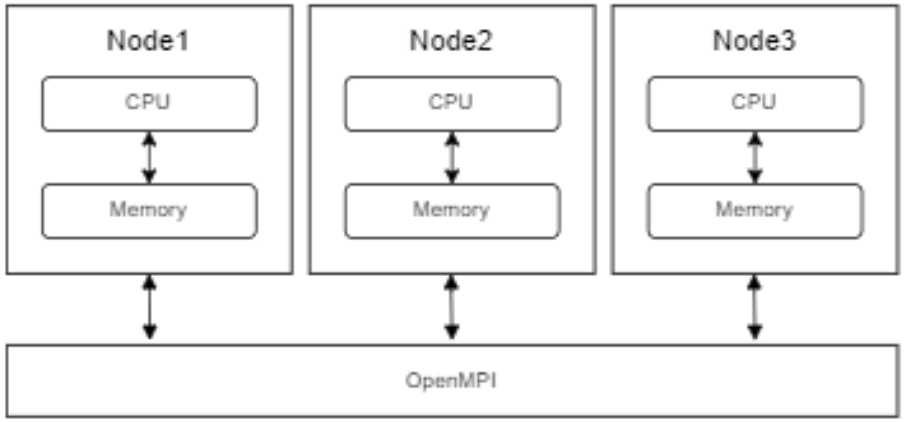
- Mendel Linux
- Lokaal netwerk
- NFS (Ganesha)
- Open MPI
- Beheerders script

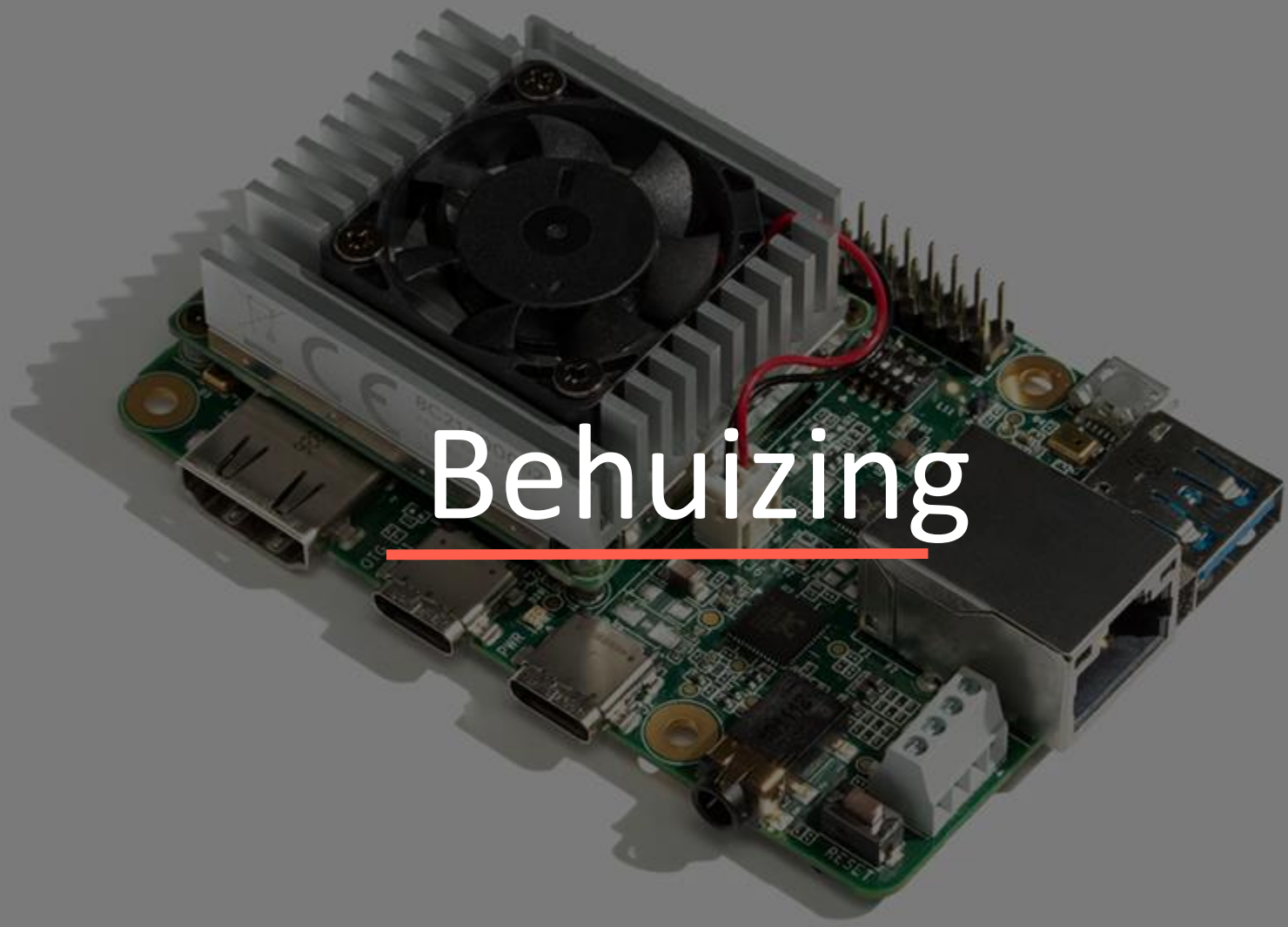


OpenMPI

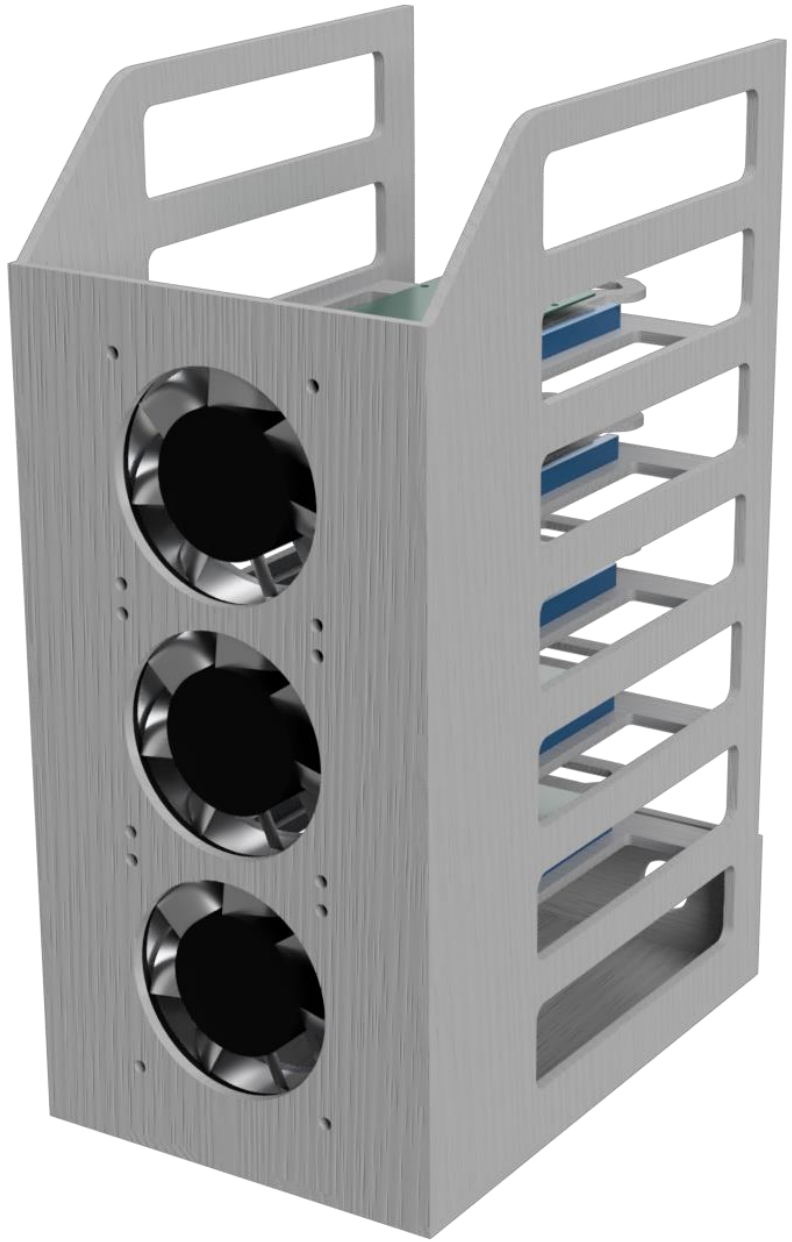


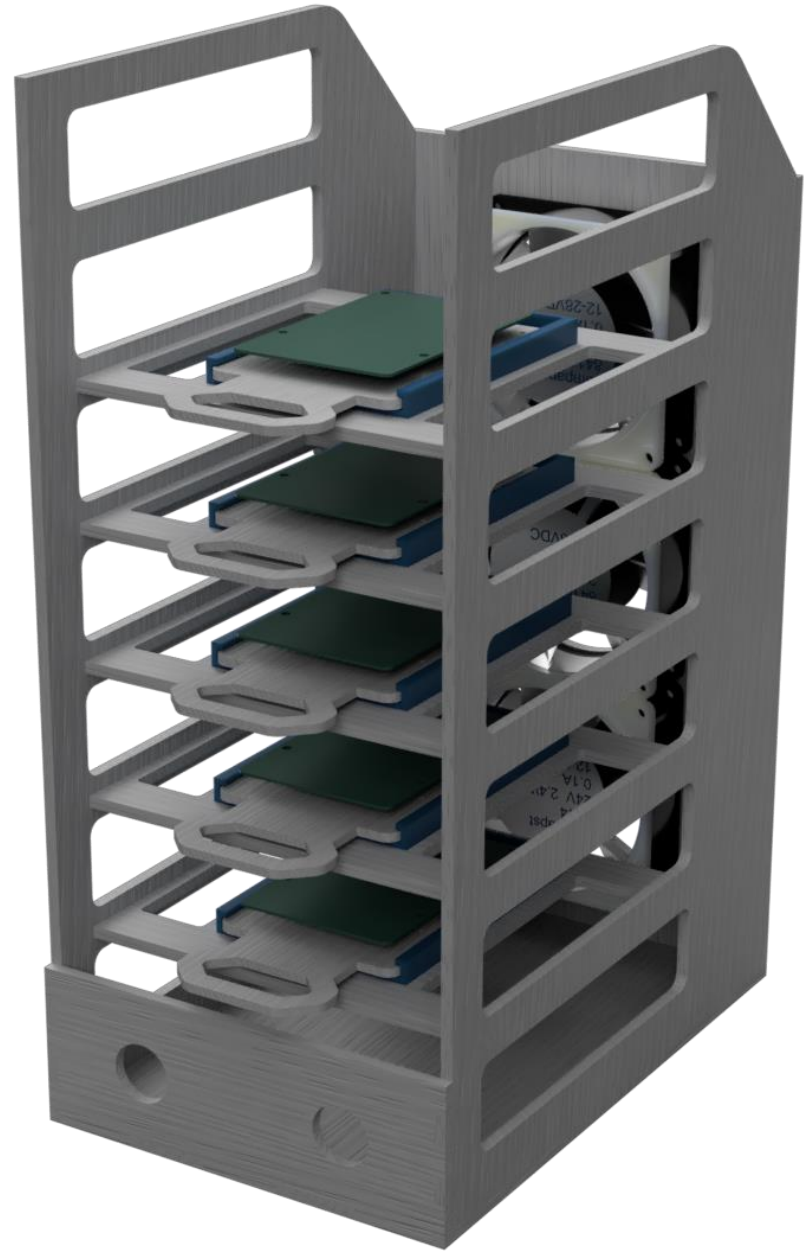
- Verdeling van taken
- Uitvoering van de taken door de nodes
- Resultaat terug naar de hoofd node sturen
- Gebruik van memory cores





Behuizing

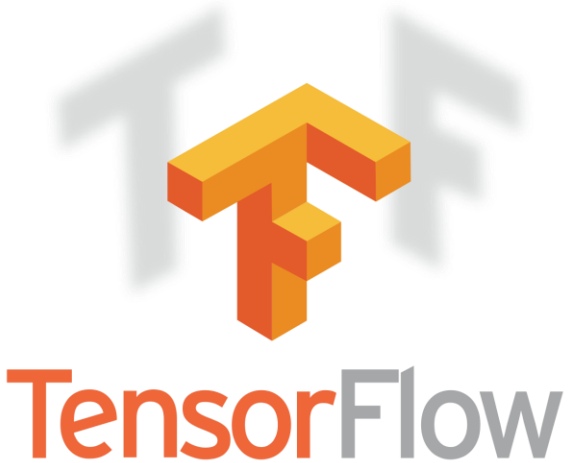




Alternatieven

Alternatieven

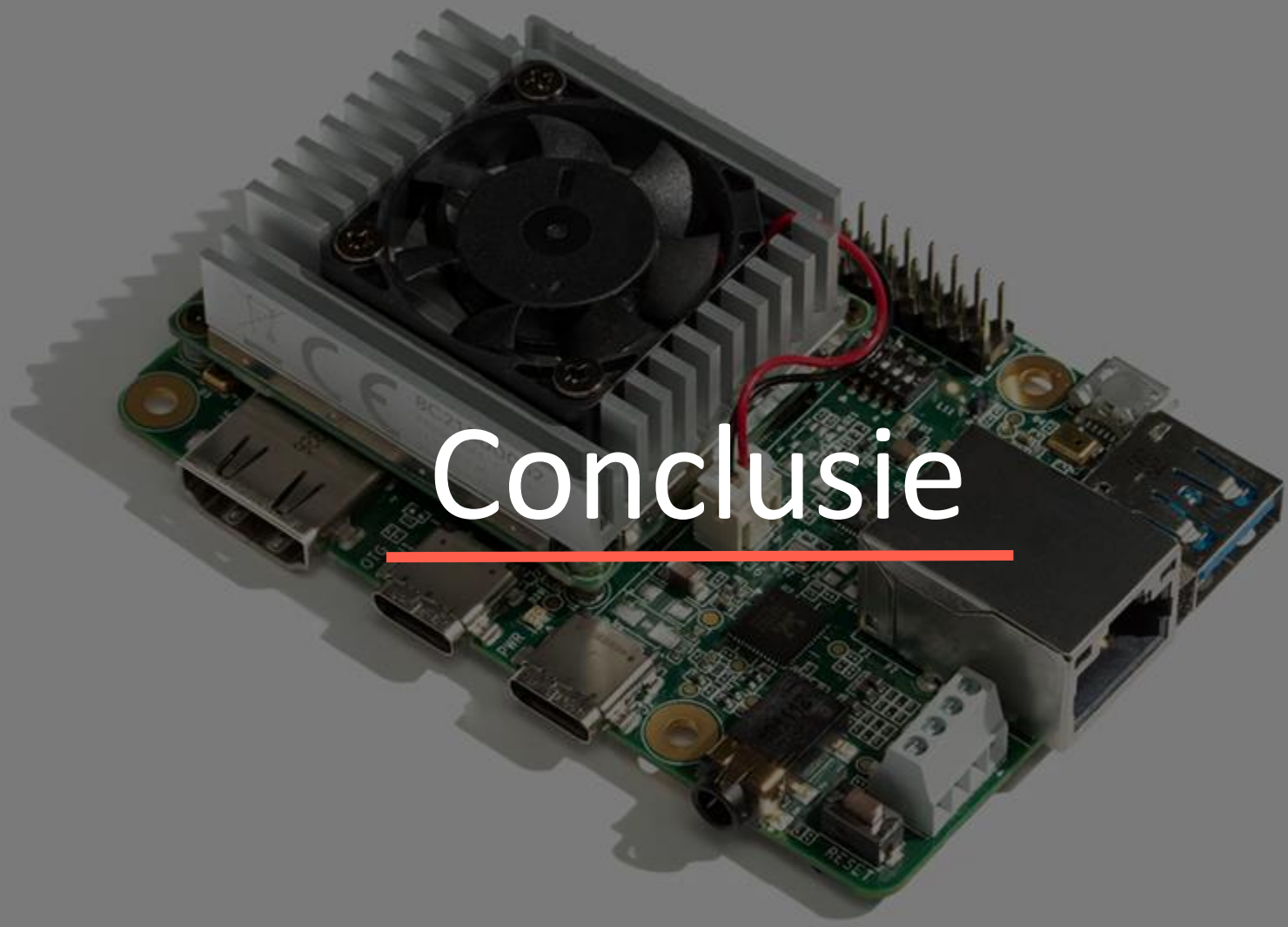
- Gebruik van TensorFlow Serving
- Afstappen van Edge TPU Dev Board cluster



Alternatieven – TensorFlow Serving

Zoals in Google datacenters

- Can serve multiple models, or multiple versions of the same model
- **Exposes both gRPC as well as HTTP inference endpoints**
- Allows deployment of new model versions without changing any client code
- Supports canarying new versions and A/B testing experimental models
- Adds minimal latency to inference time due to efficient, low-overhead implementation
- Features a scheduler that groups individual inference requests into batches for joint execution on GPU, with configurable latency controls
- Supports many servables: Tensorflow models, embeddings, vocabularies, feature transformations and even non-Tensorflow-based machine learning models



Conclusie

Conclusie

- Cluster van Coral Dev Boards moeilijk
- Vrijwel geen software ondersteuning
- Meerdere TPU modules op één "*normale*" computer
  - Goedkoper
  - Eenvoudiger
  - Bestaande software ondersteuning

PCIe x 16
Gold Finger

PCIe gen3 1 x 16

ATX POWER R/A
2x2 CONNECTOR
(spare)

FAN 4pin x 2
Header

Temperature
Sensor
EMC1422-1-
ACZL-TR

PCIe Switch 48 lane PM8533B-F3EI

Stack3
(Downstream)
PCIe lane config
8 x 2

Stack1
(Upstream)
PCIe lane
config 1 x 16

Thermal
Diode

Stack4
(Downstream)
PCIe lane config
8 x 2

RSTB

QSPI

SPI NOR FLASH
128Mb

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

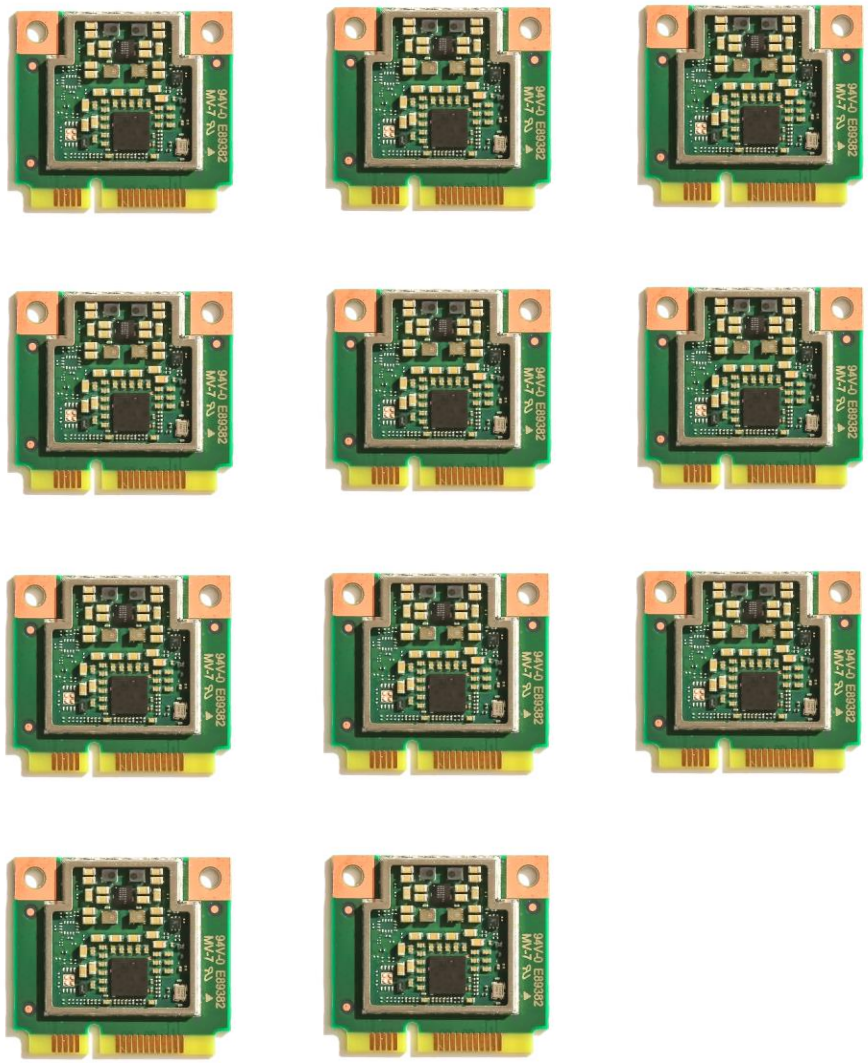
PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

PCIe Gen3 1x2 → M.2 KEY-E → M.2 TPU module

= 8x TPU

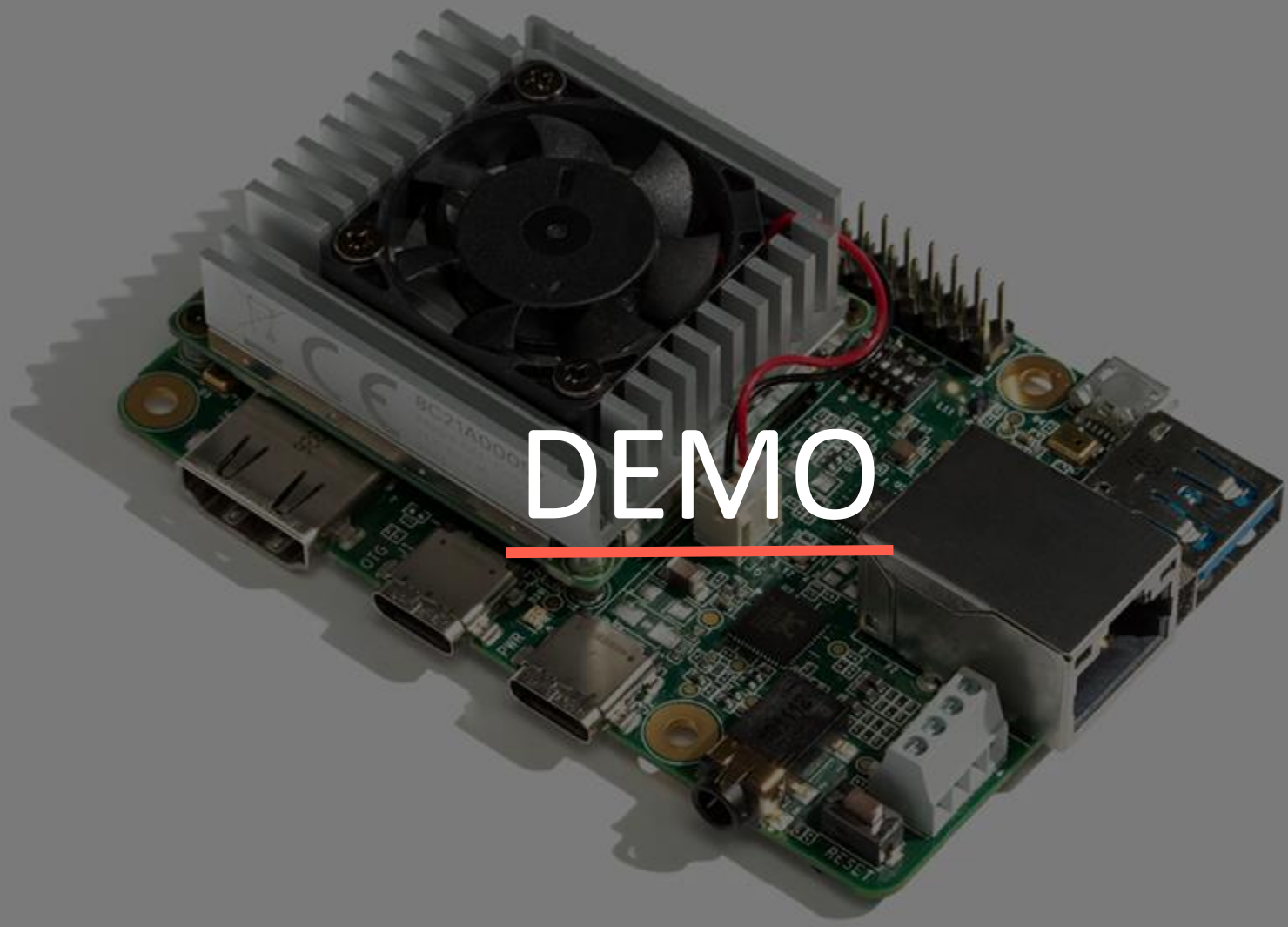


Mini PCIe Accelerators

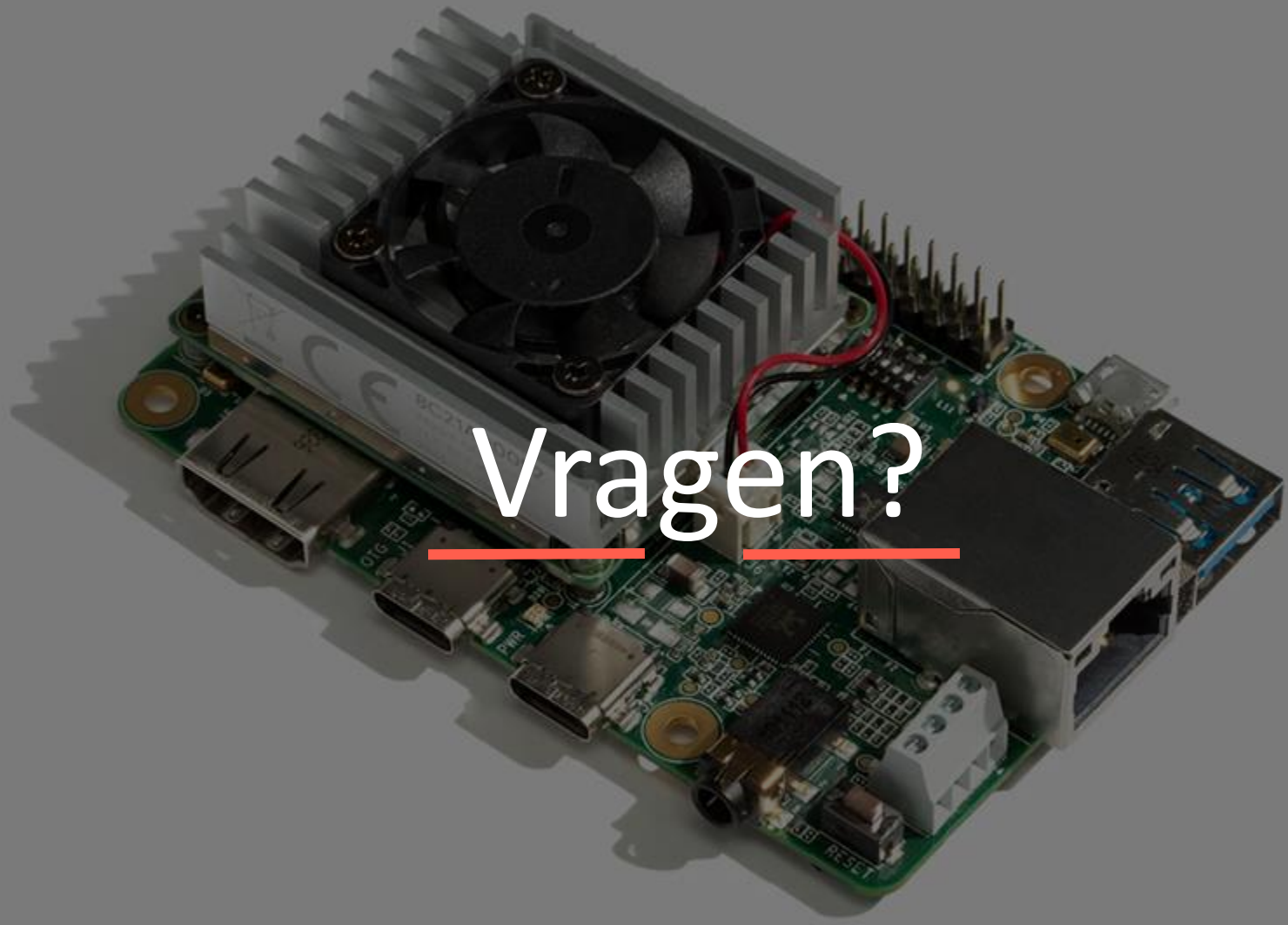
+

Biostar TB360-BTC PRO motherboard

= 11x TPU



DEMO



Vragen?